Large-Scale Machine Learning for Malware Characterization using Graphs

CYBER 20/20

Deep Machine Learning Meets Cybersecurity

John Cavazos, Ph.D.
CEO and Founder of Cyber 20/20

- Current Associate Professor at the University of Delaware
- Former JP Morgan Faculty Fellow at Inst. For Financial Services Analytics
- 20 years experience applying machine learning to problems involving code
- Spent one year with High Performance Computing Group at JP Morgan

The Problem

- Over 100K malware variants created every hour
- Bad actors have embraced automation
- Good actors still construct security defenses manually
  - Slow and error prone

The Problem

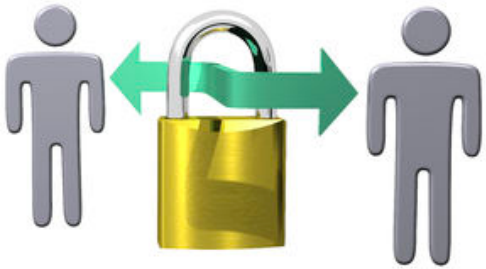
- "Mis-match" between bad actor techniques and good actor tools has caused high profile breaches to occur:

Today's Methods of Protection

Security Products

Software Patching

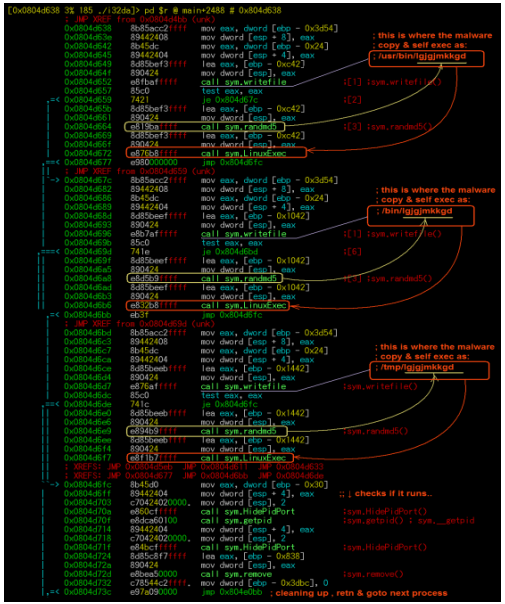
Employee Training



Secure Communication

But are these effective? Results say…not really!

Advanced Solutions Needed!

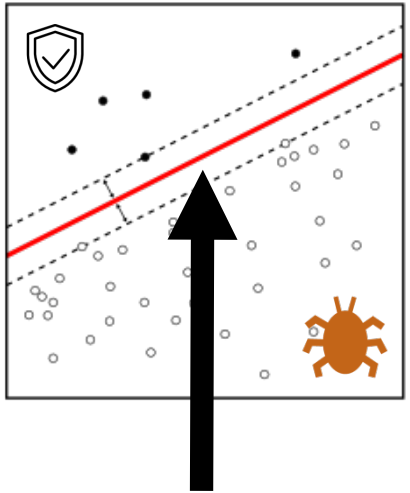


Machine Learning + Malware Characterization

What is Machine Learning?

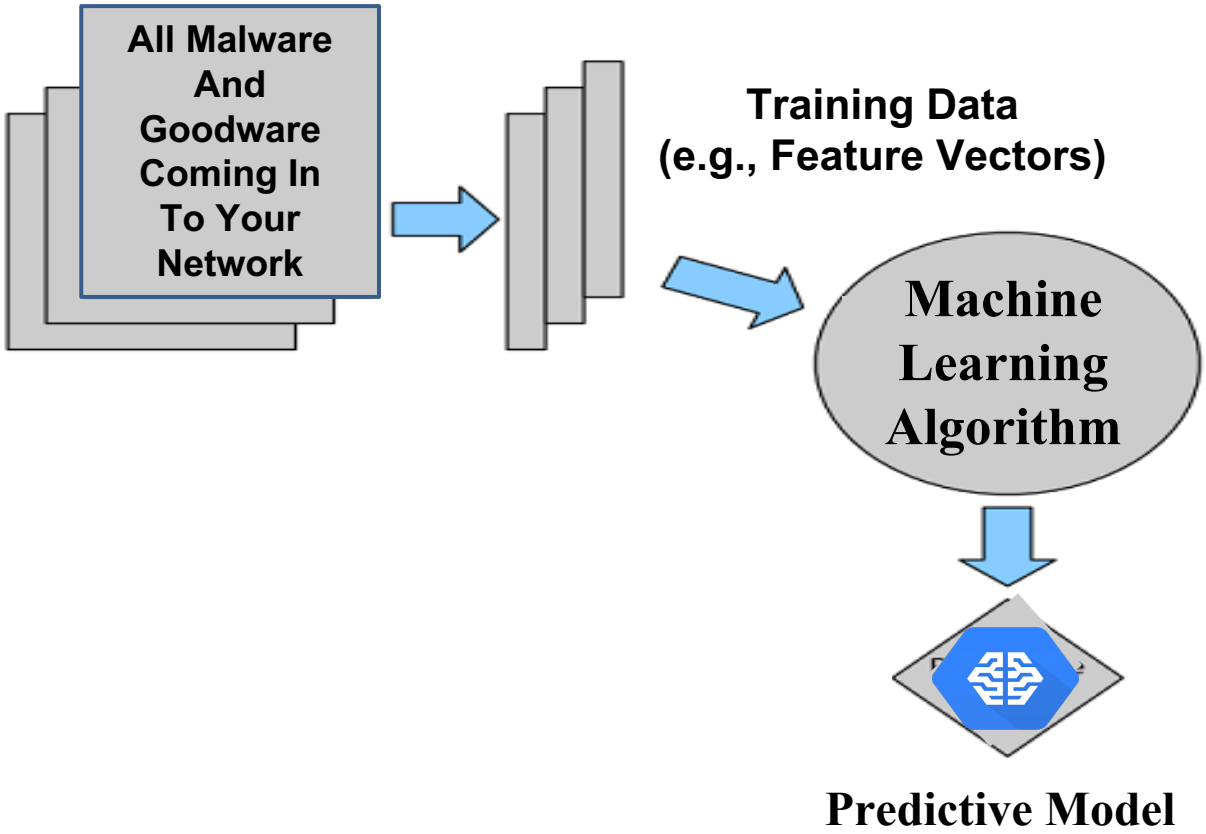
Computer algorithms that learn from and make predictions on data.

https://en.wikipedia.org/wiki/Machine_learning



Learn a good separation between Malware and Goodware

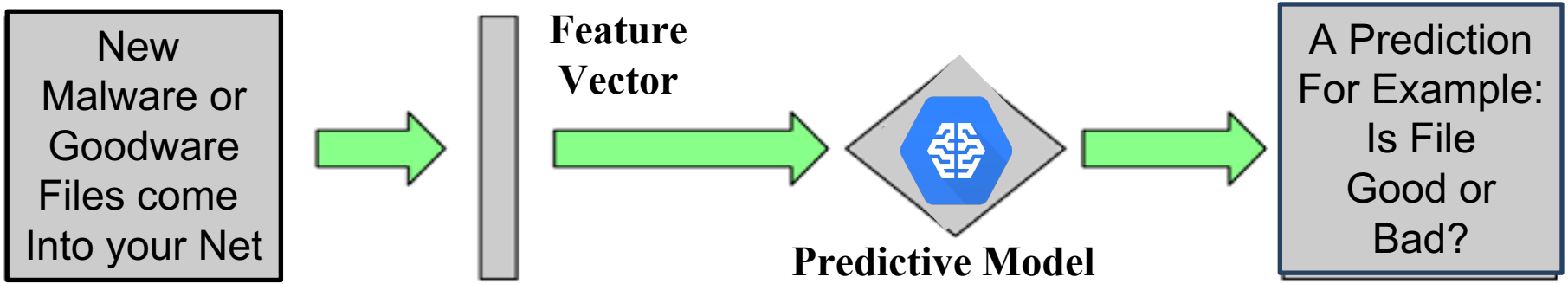
How Does Machine Learning Work?



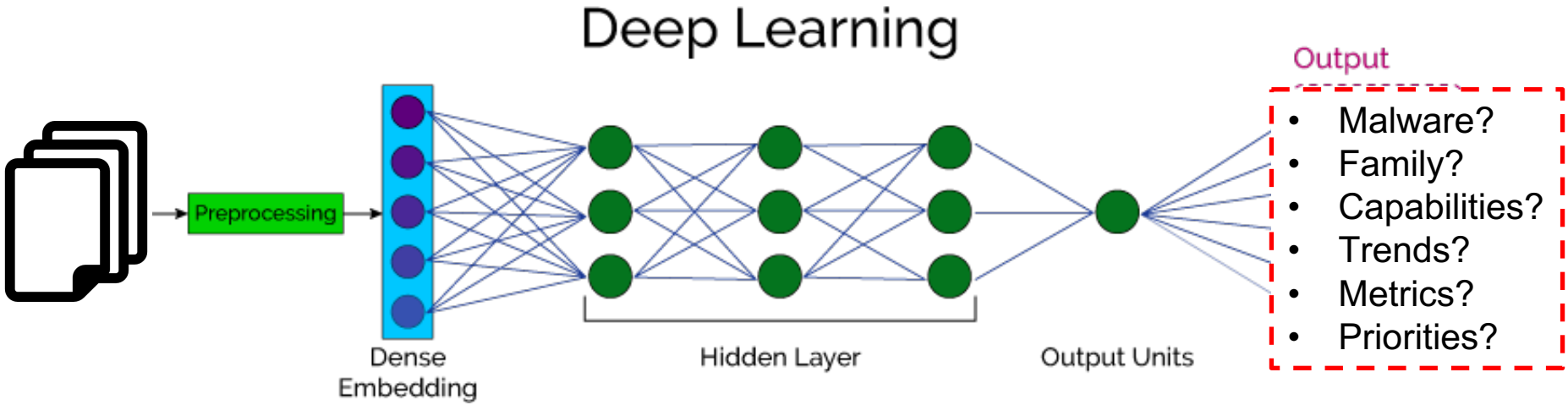
Step 1: Train Predictive Model

How Does Machine Learning Work?

Step 2: Deploy Predictive Model



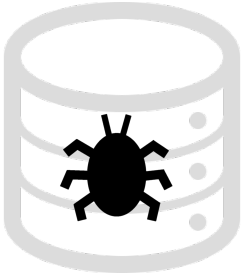
Deep Learning is a subset of Machine Learning



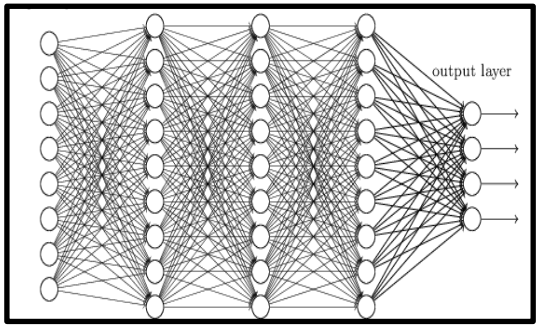
Deep Learning is a neural network with many layers of nonlinear processing units Each successive layer uses the output from the previous layer as input

The Solution: Deep Learning + Security

Why Now? Because these are readily available!



Repository of Billions of Malware
Big Data



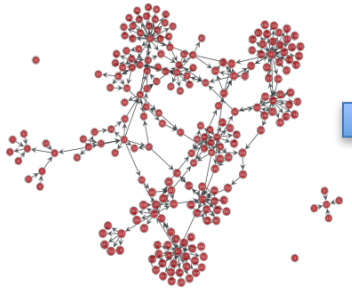
Deep Learning
Open Source Frameworks

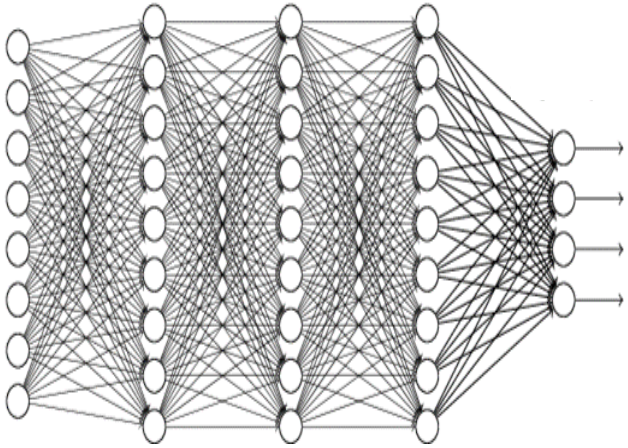
amazon web services

High-Performance Cloud Computing
Inexpensive

How? Deep Learning in the Cloud

Input





Output

- **Malware?**
- **What Family?**
- **Capabilities?**
- **Trends?**

Suspicious File

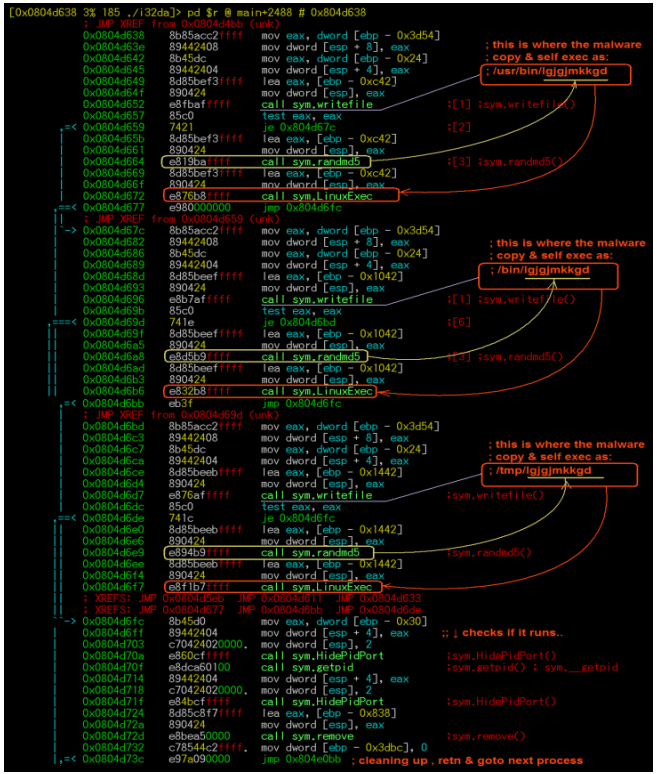
Extract Features

Cloud-Based Deep Learning Neural Network

Predictions

Deep Learning models can predict malware with very high accuracy and at network speed

What is Malware Characterization?



- Example of Malware Characteristics:
  - Is it file encrypted?
  - Does it send spam?
  - Does it read / steal private information?
  - Will it encrypt data?
  - Does it make a network connection?
  - What does the reverse-engineered code look like?

Malware Characterization

Bytes

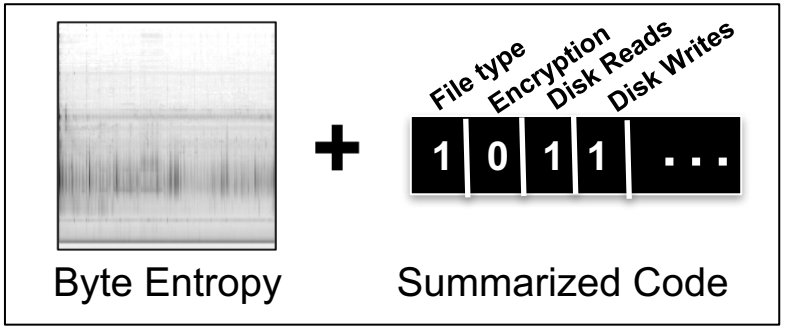
- Shannon Entropy
- Bytes N-grams
- Strings

Code

- Instruction N-grams
- Statistics
  - Function, Blocks
  - Calls, Branches

Malware Characterization

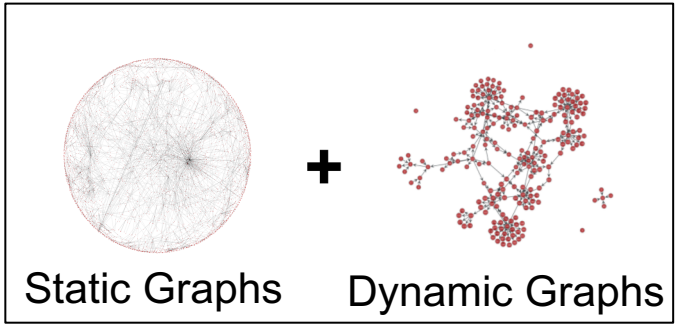
State-of-the-Art



- Feature Vector Characterization
  - Byte Entropy Histograms
  - Summarized Code Structure

vs

Graph-Based



- Graph-Based Characterization
  - Static Code Graphs
  - Dynamic Behavioral Graphs

Malware Classification using Machine Learning

Malware Stream

Data from Reversing Labs

- Billions of malware
- Curated streams

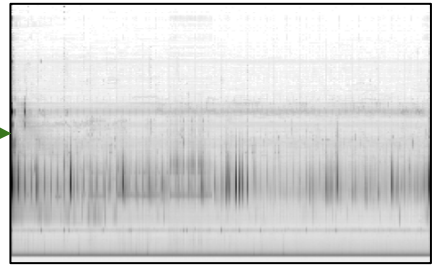
Financial stream

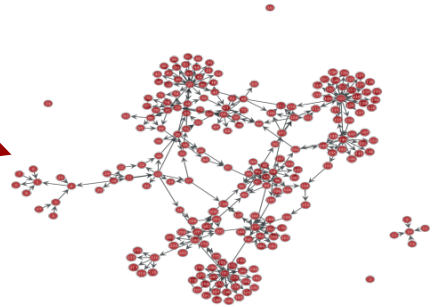
- 1.65 millions
- 40+ families

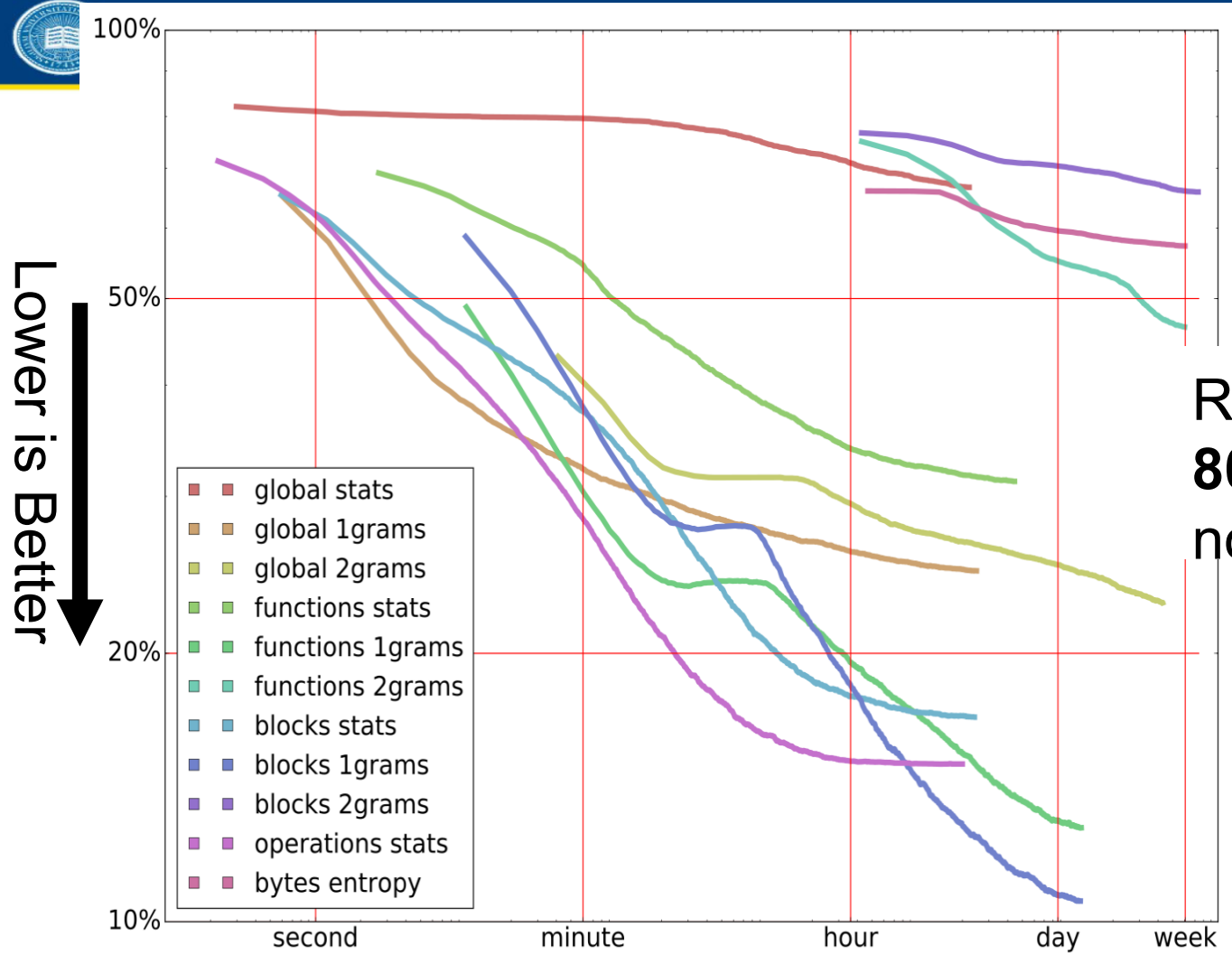
Selected 11 families with > 1000 malware

Malware Features for Deep Learning

	Characterization		Format	Size
State-of-the-Art	bytes-entropy histogram		matrix	256 x 256
	Global	statistics	vector	43
		1-grams		53
		2-grams		2809
Graph-Based	Function	statistics	matrix	20 x 23
		1-grams		20 x 53
		2-grams		20 x 2809
	Block	statistics	matrix	20 x 10
		1-grams		20 x 53
		2-grams		20 x 2809
	Operations	statistic	matrix	20 x 2

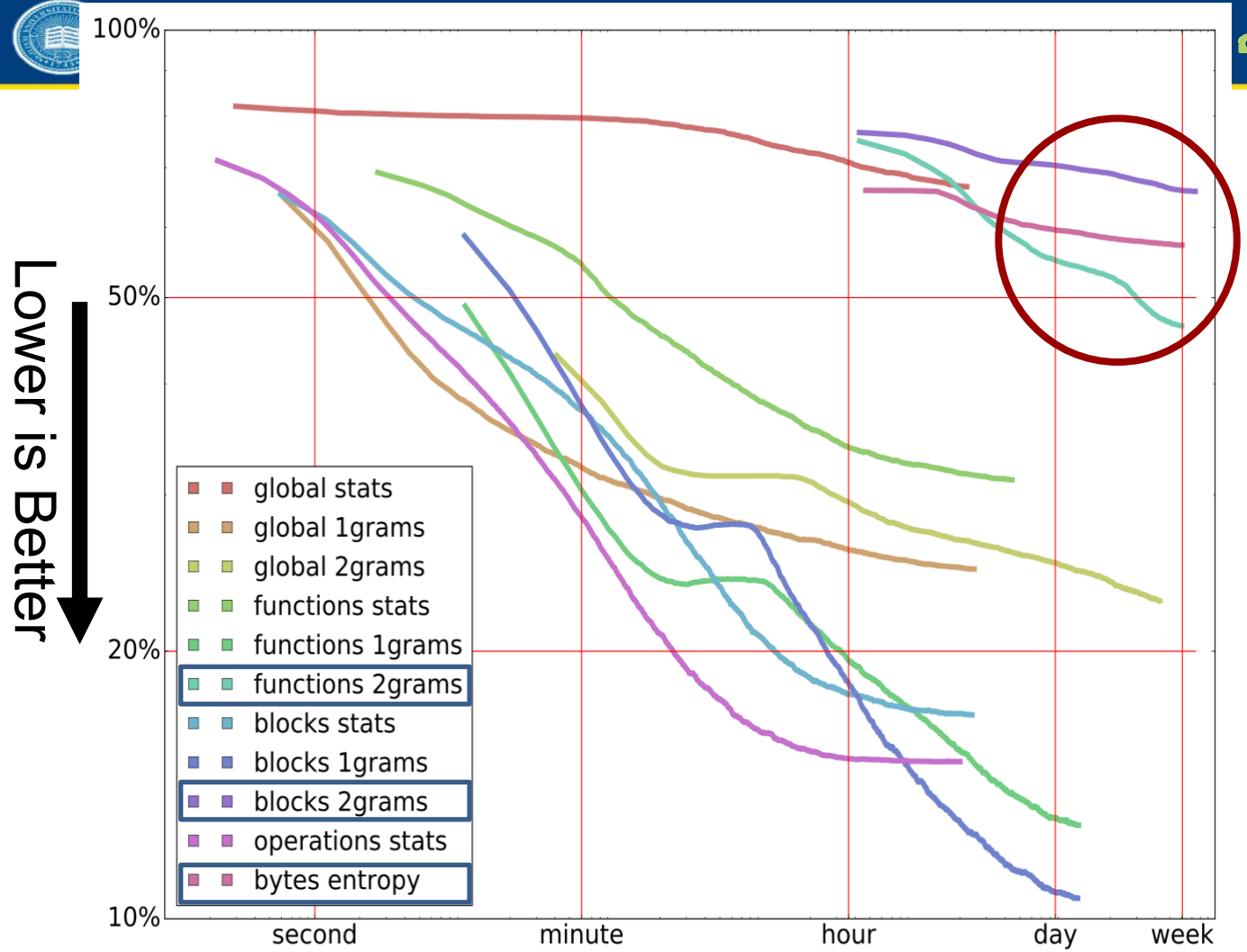






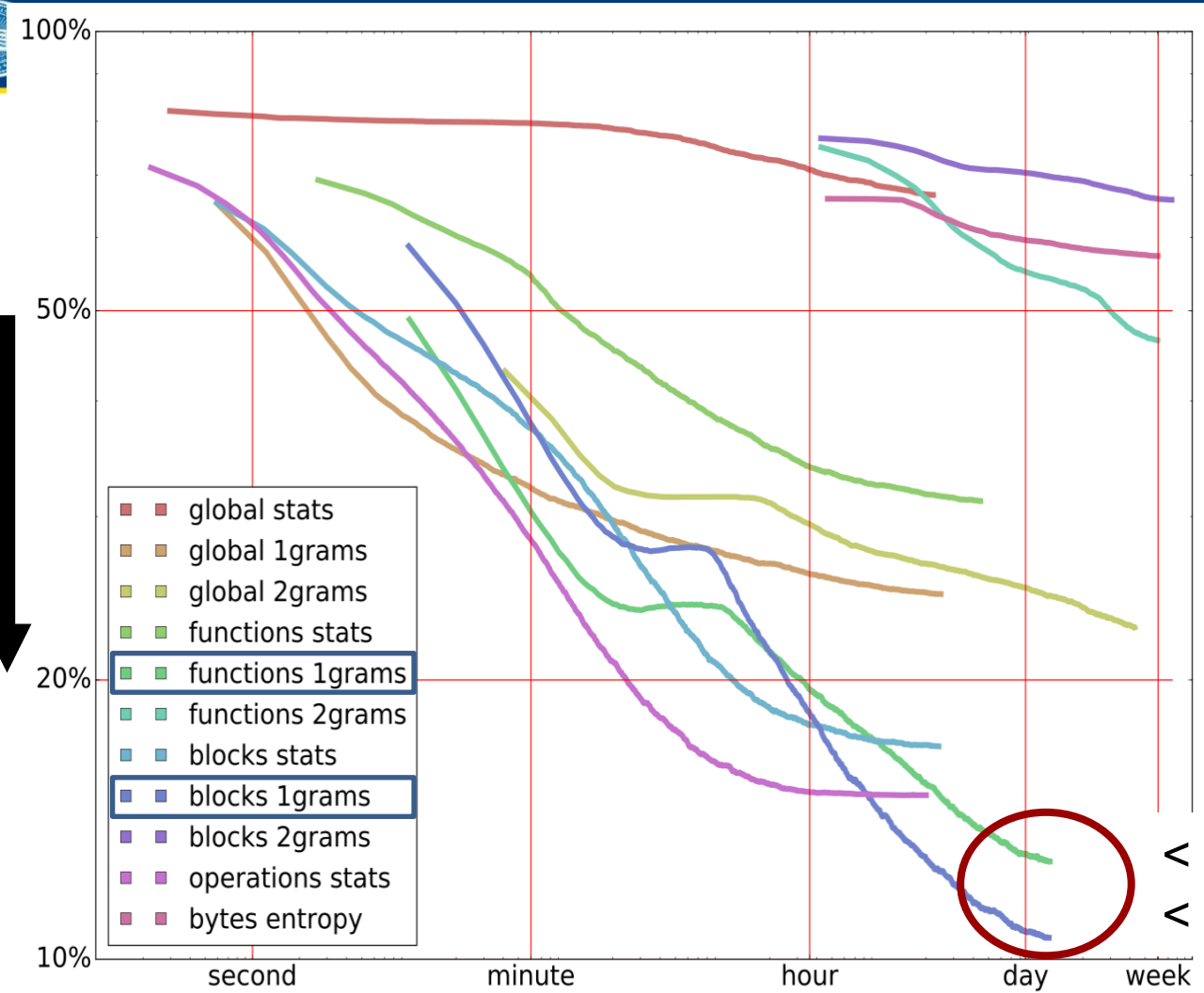
Lower is Better

Results running on **8000** compute nodes for **2 weeks**



Characterization too large

- Blocks 2grams
  Size: 20 x 2809
- Byte entropy
  Size: 256 x 256
- Functions 2grams
  Size: 20 x 2809



Lower is Better

Graph-Based

- Functions 1grams
  Size: 20 x 53
- Blocks 1grams
  Size: 20 x 53

< 13%

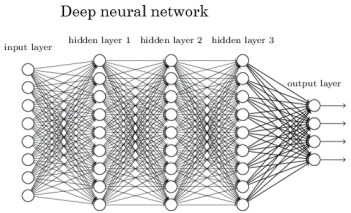
< 11%

Summary of Best Results: Ensemble of DNNs

Models	1 best	2 best	5 best
BEH & Global level	16.0%	15.3%	13.8%
Graph-Based	8.4%	8.0%	8.0%
All Features	6.9%	6.5%	6.3%

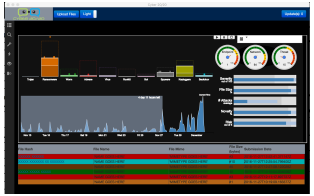
50% improvement in accuracy using Graphs!

An all-software product consisting of:



1) a deep learning neural network in the cloud

2) a network tap ("threat analytics platform") used to extract and examine all files from traffic entering a critical infrastructure network;



3) An analytics user interface to view results and provide visibility;

4) Optional: Threat Intelligence feed into deep learning "engine."

Questions ?

Deep Machine Learning
Meets Cybersecurity

John Cavazos
CEO and Founder
cavazos@cyber2020.com
302-690-6041

WWW.CYBER2020.COM